RobotReviewer: Evaluation of a System for Automatically Assessing Bias in Clinical Trials

Iain J Marshall, Department of Primary Care and Public Health Sciences, King's College London, UK

Joël Kuiper, University Medical Center, University of Groningen, Groningen, The Netherlands

Byron C Wallace, School of Information, University of Texas at Austin, Austin, Texas, USA

Correspondence to: iain.marshall@kcl.ac.uk

Abstract

Objective To develop and evaluate RobotReviewer, a machine learning (ML) system that automatically assesses bias in clinical trials. From a (PDF-formatted) trial report, the system should determine risks of bias for the domains defined by the Cochrane Risk of Bias (RoB) tool, and extract supporting text for these judgments.

Methods We (pseudo-)annotated 12,808 trials using data from the Cochrane Database of Systematic Reviews (CDSR). Trials were labeled as being at low or high/unclear risk of bias for each domain, and sentences were labeled as being informative or not. This dataset was used to train a novel multi-task ML model. We estimated the accuracy of ML judgments versus humans by comparing trials with two or more independent RoB assessments in the CDSR. Twenty blinded experienced reviewers rated the relevance of supporting text, comparing ML output with equivalent (human-extracted) text from the CDSR.

Results By retrieving the top 3 candidate sentences per document (top-3 recall), the best ML text was rated more relevant than text from the CDSR, but not significantly (60.4% ML text rated 'highly relevant' *v* 56.5% of text from reviews; difference +3.9%, [-3.2% to +10.9%]). Model RoB judgments were less accurate than those from published reviews, though the difference was <10% (overall accuracy 71.0% with ML *v* 78.3% with CDSR).

Conclusion Risk of bias assessment may be automated with reasonable accuracy. Automatically identified text supporting bias assessment is of equal quality to the manually identified text in the CDSR. This technology could substantially reduce reviewer workload and expedite evidence syntheses.

Background and significance

Assessing bias is a core part of systematic review methodology, and reviews typically use standardized checklists or tools to assess trial quality.(1) The Cochrane Risk of Bias tool is one such tool.(2) It has been adopted across the Cochrane Library, and increasingly in systematic reviews published elsewhere. The tool comprises seven core domains (see Box 1), which reviewers score as being at *high*, *low*, or *unclear* risk of bias.

The risk of the following types of bias is assessed as being *high*, *low*, or *unclear*:

- Random sequence generation
- Allocation concealment
- Blinding of participants and personnel
- Blinding of outcome assessment
- Incomplete outcome data
- Selective outcome reporting
- Other

For the purposes of this study, the domain 'Other' is not considered further, as it is idiosyncratically applied across different reviews.

Box 1 Items from the Cochrane Risk of Bias

Bias assessment is time-consuming: it takes experienced reviewers around 20 minutes for every study included in a systematic review.(3) The requirement to assess risks of bias has been identified as an important factor preventing Cochrane reviews from being kept up to date.(4) Bias assessment is also subjective: individual reviewers have been found to have low rates of agreement,(3) though this improves somewhat when review specific guidance is provided.(5)

Technology to assist reviewers in assessing bias thus has the potential to substantially reduce workload. An accurate system could make bias assessment quicker and more reliable, thereby freeing up researcher time to concentrate on thoughtful evidence synthesis. Ultimately, more efficient bias assessment would help keep systematic reviews up to date.(6)

In this paper, we introduce a novel machine-learning (ML) approach, which models risks of bias simultaneously across all domains while identifying text supporting these judgments. We evaluate this algorithm against risk of bias assessments from published systematic reviews, and by using the expert assessment of 20 experienced systematic reviewers.

Our preliminary work demonstrated the feasibility of automated risk of bias assessment.(7) Here we extend our model and present an evaluation which allows assessment of whether the technology is mature enough to be used in practice.

Objectives

We describe the development and evaluation of RobotReviewer, a system to automate the assessment of bias of randomized controlled trials using the Cochrane Risk of Bias (RoB) Tool. For each domain in the Cochrane RoB tool, the system should reliably perform two tasks: 1. Determine whether a trial is at low risk of bias (document classification of low *v* high or unclear risk of bias), and 2. Identify text from the trial report that supports these bias judgments (sentence classification of relevant *v* irrelevant). In the evaluation, we aim to compare model performance with the quality of Risk of Bias assessments in published systematic reviews, to help judge to what extent bias assessments could be automated in real review production.

Methods

Most recent approaches for automating biomedical data exaction have used *supervised learning*, in which algorithms learn from manually annotated documents.(8) These are promising approaches, but collecting human annotations is time-consuming and therefore expensive. This is especially true for biomedical text mining tasks (such as assessing risks of bias), as these require expert and therefore costly annotators. Because of this, corpora used on similar tasks to date are relatively small, comprising 100-350 documents.(8,9)

Here we take a different approach: to obtain sufficient labeled data we use *distant supervision*,

an emerging machine learning methodology that exploits structured data in existing databases in place of direct human supervision. For example, Mintz *et al.* used the online Freebase encyclopedia to recognize relations between named entities in natural language.(10) And more similar to our work here, Ling *et al.* used extant resources to train models for generating gene summaries.(11)

Distant supervision involves deriving labels for unlabeled data using structured resources. This derivation is typically done according to some rules or heuristics that cover the majority of cases but that are imperfect (e.g., string matching). This produces *noisy* labels (having a higher rate of errors than manual annotation). However, because these are 'free' labels, we can build and exploit larger training datasets than would be otherwise feasible. Larger training datasets, in turn, have been shown to improve model performance.(12)

domain:	Allocation concealment
risk of bias:	High
justification:	Quote: "…using a table of random numbers." Comment: Probably not done.

Box 2 Example of the risk of bias data stored in a Cochrane review for the domain allocation concealment, from (12)

Here we derive distant supervision from the Cochrane Database of Systematic Reviews (CDSR). The CDSR comprises more than 5,400 systematic reviews on health produced by members of the international non-profit organization, the Cochrane Collaboration. This dataset includes expert risk of bias assessments for clinical trials (an example is presented in Box 1). Crucially, in a substantial minority of these assessments, the review authors include direct quotes from the original trial report to justify their judgments. Therefore, in addition to the *article-level* risk of bias labels, we can also derive sentence level annotations within full-texts that indicate whether a given sentence was used in assessing the risk of bias for a particular domain. This derivation is accomplished by string matching. Figure 1 illustrates this schematically.

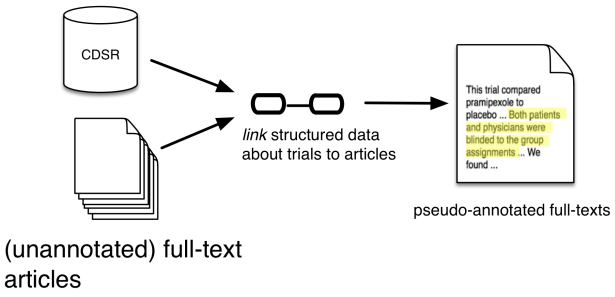

Figure 1: Algorithmic annotation of clinical trial PDFs using data from the CDSR

Automating the labeling of the clinical trials corpus

Our method for corpus construction is outlined in Figure 2.

Trial linkage

First, we sought full-text PDFs of clinical trials that were included in systematic reviews in the CDSR. The CDSR contains semi-structured reference data, but not unique identifiers. We therefore used the following strategy, designed for high precision. For each trial included in a systematic review in the CDSR we conducted multiple searches of PubMed. Each search used non-overlapping subsets of citation information, any of which might be expected to uniquely retrieve the trial (e.g. search 1: articles matching full title; search 2: articles with exact author combination with matching publication year; search 3: articles with matching journal name, volume, issue, and page number). We considered a positive match where two or more searches retrieved the same article. We were able to link 52,454 out of 67,894 studies included in systematic reviews in the CDSR to a unique publication using this method, and obtained 12,808 of these publications in PDF format.

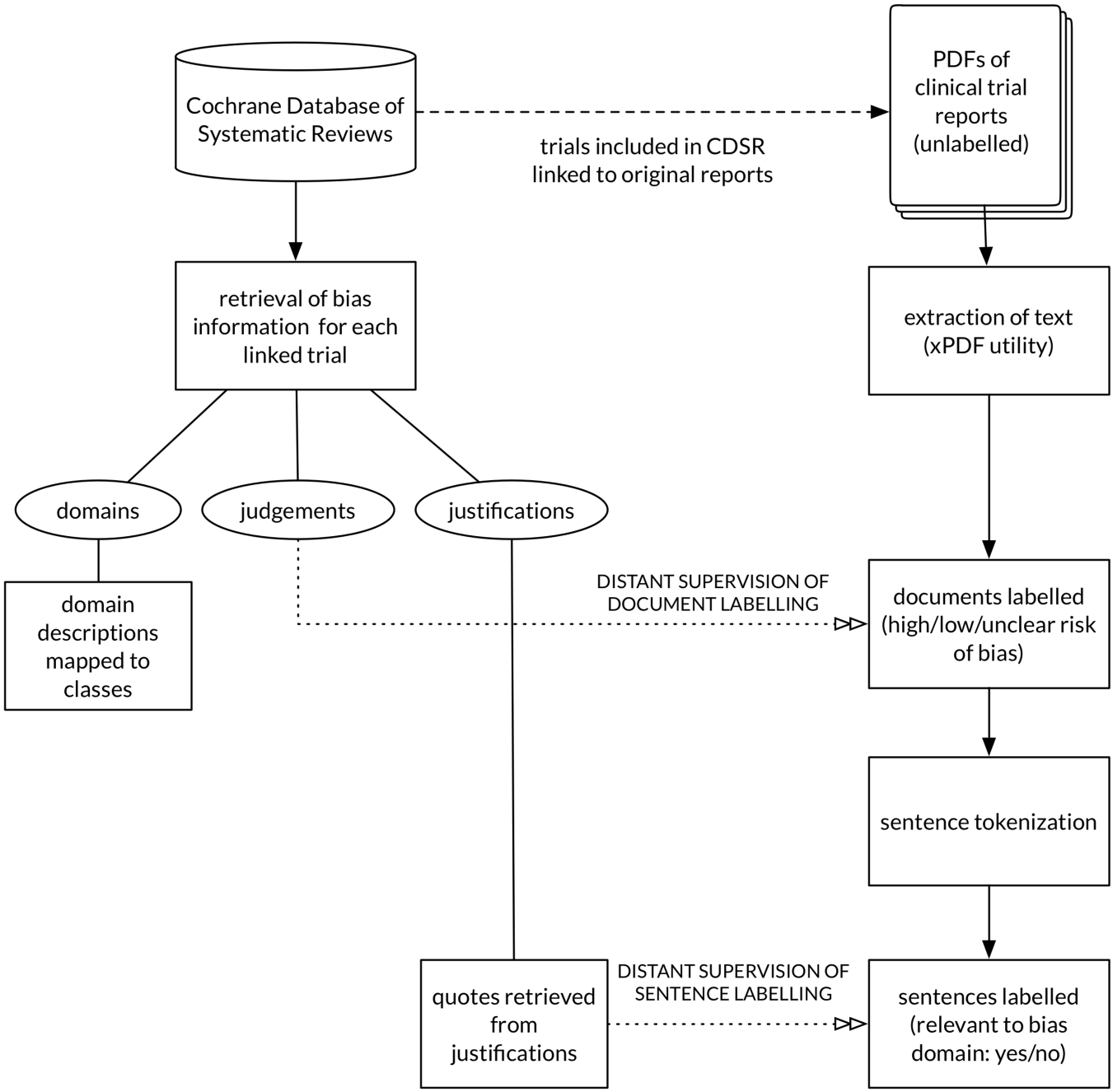

Figure 2: Schematic of corpus construction, and outline of the distant supervision process

Pre-processing of CDSR data

Although the Risk of Bias tool assesses seven core outcomes, in practice there is much variation in the approaches, and Cochrane review authors are free to choose domains of bias they feel most relevant, or even add their own. Across the whole of the CDSR, there are more than 1400 unique strings identifying bias domains. Most of these referred to one of the seven core domains listed in Box 1. We manually mapped these alternate descriptions to the domain label, and excluded other areas of bias idiosyncratic to individual reviews. For each linked study, we extracted the types of bias assessed, the bias judgments, and the justifications for the judgments (see Box 2).

Labeling PDFs using distant supervision

Plain text was extracted from the PDFs using the pdftotext utility from xPDF.(13) The extracted text was tokenized into sentences and words using a Punkt tokenizer adapted for scientific research.

For task 1 (document annotation), we algorithmically labeled each document as being at 'low' or 'high/unclear' risk of bias, using the judgment from the linked Cochrane review. We dichotomized this outcome both because it enabled the use of a binary model, and because it fits with the typical conduct of a systematic review (sensitivity analyses are frequently conducted including only studies at low risk of bias).

For task 2 (sentence annotation) where quote data was available in the linked Cochrane review, sentences containing exactly matching text were labeled as relevant to the risk of bias domain. All other sentences were labeled as irrelevant.

This strategy produces incomplete labels and specifically would be expected to have high precision and low recall. Cochrane review authors are likely to quote one or two sentences that they feel best justify a risk of bias judgment. Ideally, all text relevant to the bias decision would be labeled.

Machine Learning Approach

We have developed a novel *multi-task* machine learning model that maps articles to risk of bias assessments (low or high/unclear) (task 1) and simultaneously extracts sentences supporting these judgments (task 2). Multi-task learning refers to scenarios in which we are to induce classifiers for multiple, related classification problems or 'tasks' (14). In our case, bias assessment for the respective domains constitute our related tasks.

Our approach includes two novel components. First, we explicitly incorporate features derived from sentences that support risk of bias assessment into the 'document level' model that predicts the risk of bias. Second, we jointly model risk of bias across all domains of interest, for both sentence and document level predictions.

To incorporate supporting sentence information into the document level model, we introduce features that encode specific tokens appearing in the sentence(s) deemed to support the risk of bias assessments for the respective domains. We have presented technical details of this method elsewhere (7), but briefly summarize the approach here for completeness. Intuitively, if the word 'computer' appears in the sentence that supports the assessment for the *randomization* domain for a given article, it is a strong indicator that this article has a low risk of bias with respect to randomization. The features we introduce encode such information. At test time, however, we will not know which sentences support which risk of bias judgments (the majority of sentences will not be relevant for RoB assessment). Therefore, prior to document-level risk of bias prediction we automatically generate special indicator features that encode all tokens extracted from sentences deemed by the sentence-level model as likely to support RoB assessments. In summary: when training models, we capitalize on features that encode tokens comprising sentences crossed with binary indicators that these sentences were used to support risk of bias assessment. At test time, we replace the indicators with *predictions* for each domain regarding sentence relevance.

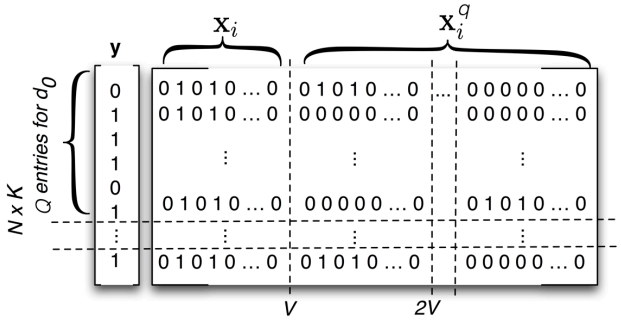

Figure 3: A schematic depiction of our multi-task learning approach. We define a joint classification

model across domains. To achieve this, we include k representations of each instance (e.g., document) d in the design matrix, one per risk of bias domain. We construct the target vector y with the corresponding per-domain labels for d (derived from the CDSR). Each of the k entries for d comprises k+1 concatenated vectors of length equal to the vocabulary size (V), where for the document model this vocabulary is the union of unique uni- and bi-grams appearing in at least two articles, and for the sentence model V is the union of uni- and bi-grams appearing in at least two supporting sentences. The first copy of V in each row is shared across all entries representing the corresponding article; the remaining are domain specific copies. Thus in any given row, all but two of these sub-vectors will be zero vectors. Specifically, each row i will contain two non-zero vectors that are copies of the bag-of-words representation (binary indicators for V) for instance i (x_i): one shared and the other domain specific. The shared component allows borrowing of strength across domains, while the domain-specific components enable the model to learn words that signal low risk of bias (or the likelihood of supporting RoB assessment, for the sentence prediction task) only in specific domains

Here we extend this model to borrow strength across risk of bias domains, using a *multi-task* approach for both sentence and document level classifications. More specifically, we introduce 'interaction features' that represent the intersection of domains and token (word) indicators. Denoting the number of domains by *k*, we insert *k* copies of each feature vector **x** (one per domain) for each instance, in addition to a shared copy of **x** common to all domains (see Figure 3 for a schematic of this approach). For example, there will be a feature corresponding to the presence of the word 'computer' *and* the target domain *randomization*. This will be non-zero only in columns that comprise the copy of **x** specific to *randomization*. Note that this can be viewed as an instantiation of Daumé's *frustratingly easy* domain adaptation approach.(16) Our sentence model is thus trained jointly across all risk of bias domains; the shared component enables information sharing between them. We adopt this approach for both the sentence and the document level models.

For a new article (seen at test time), we then use the multi-task sentence model to generate predictions for each sentence regarding whether it is likely to support assessments for the respective domains. Before a document level prediction is made regarding the risk of bias, indicators corresponding to the tokens comprising sentences predicted to be relevant are inserted into the vectors representing documents. This is done for each domain. Note also that these document representations also include a shared component across domains (again enabling borrowing of strength). Therefore, sentence level predictions (made via a multi-task sentence-level model) directly inform our multi-task document level model. This realizes a joint approach to predicting sentence level relevance and document level assessments across related tasks.

For both the sentence and document level model we adopt a linear classification model defined by a weight vector **w**, such that $y = L(\mathbf{wx})$, where L maps the continuous score to a categorical label of 0 (when $\mathbf{wx} < 0$) or 1 (when $\mathbf{wx} \geq 0$). We use the *hinge-loss* function, which imposes no loss when the model prediction is correct and a loss proportional to the magnitude of **wx** when the prediction is incorrect. We combine this with a squared L2 penalty term on the model parameters to form our objective. We fit this model (estimate **w**) to minimize our objective via Stochastic Gradient Descent (a standard optimization procedure in which one optimizes parameters to minimize an objective function by following the gradient as approximated by iterative evaluation on individual instances). We tune the parameter, trading regularization strength (model simplicity) against empirical loss via line search ranging over values from 10^{-4} to 10^{-1}, equidistant in log-space.

All models we consider leverage token-based features encoded in a binary 'bag-of-words' representation. For sentence prediction, our vocabulary *V* comprises all unique uni- and bi-grams present in at least two supporting sentences. For document prediction, *V* comprises all uni- and bi-grams that appear in at least two articles. We preprocessed the text by removing English ‘stop words’ (uninformative words like “the” and “as”) and converting all text to lowercase. Because using full-text and interaction features produces a very large feature space, we use the ‘feature hashing’ trick to keep the model tractable. The hashing trick maps strings to vector indices via a hashing function. This has been shown to work well for large-scale multi-task text classification.(17)

To summarize, we are using a novel model for risk of bias prediction that (1) jointly makes article- and sentence-level predictions, and, (2) borrows strength across the related risk of bias assessment tasks via multi-task learning, for both article and sentence predictions.

We have made the code used for the entire distant supervision pipeline and evaluation available at www.github.com/ijmarshall (under cochrane-nlp, and cochrane-nlp-experiments). End users (i.e., systematic reviewers) might instead use our prototype web-based tool, Spá (19), which generates and presents risk of bias assessments for articles uploaded by users (Figure 5).

Evaluation

Task 1: Document prediction

For task 1 (document judgments) we exploited the fact that many trials are described in more than one Cochrane review (ranging from 239 to 1148 trials with domain relevant information in both Cochrane reviews). Thus, we were able to obtain a second, independently conducted risk of bias assessment for the articles comprising this set of trials.

We used this set as our held-out test data with which to compare the relative accuracy of the automated versus manual strategies. Specifically, we used one set of human risk of bias assessments as the evaluation gold standard, and the second as a surrogate for human performance. This allowed us to establish an upper-bound on the `accuracy' we can hope to see from an automated system, given that human agreement is not perfect. Figure 4 depicts our evaluation setup schematically.

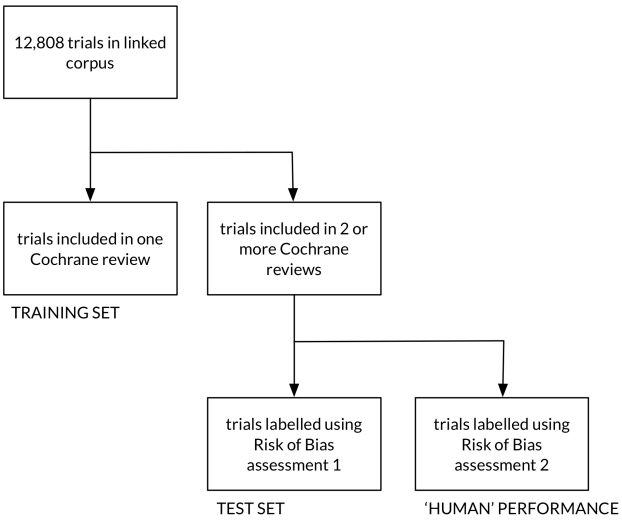

Figure 4: Test/training set partition for the document-level judgments, and use of the second Risk of Bias assessment as a surrogate for human performance

For document prediction, we consider three models. We evaluate only the binary outputs (0/1) of these predictive models (i.e., we do not consider probabilities associated with these predictions). Model 1 is a standard uni- and bi-gram model using the entire document text; each RoB domain is modeled independently in this approach.

Models 2 and 3 are multi-task models that incorporate the supporting text (output from task 2) and also share features across domains for document level prediction. Model 2 uses the multi-task approach to jointly model all 6 domains. Model 3 is otherwise identical to model 2, but excludes *incomplete reporting of outcomes* and *selective reporting*. We excluded these domains post hoc under the assumption that predictions in these noisy domains were adversely affecting performance in other domains in the case of our multi-task model. Specifically, *incomplete reporting of outcomes* would typically involve calculation of withdrawal and dropout rates (which is not possible using bag-of-words modeling), and *selective reporting* would usually require reference to a trial protocol, not available to our algorithm currently. Finally, for comparison, we report a *baseline* result, where all documents are labeled with the majority class for the domain (determined from the training set).

Task 2: Sentence prediction

For task 2, we automatically labeled held-out documents using the trained model. Documents were labeled by our model using two strategies:

top-1, where the top scoring sentence per document was identified, and top-3, where the top three scoring sentences were identified. We then used two control strategies: *cochrane*, where the text justifying a bias decision was taken directly from the published review in the CDSR describing the trial (to estimate human performance; mean 1.3 sentences per trial), and *baseline* where a sentence was drawn randomly from the document.

We recruited an expert panel of 20 systematic review authors (median published reviews per author: 19, IQR 7.5 to 51.5), all of whom had substantial experience using the Cochrane Risk of Bias tool.

Sentences identified by each strategy were presented to the expert panel members, who were blinded to the text source. The assessors were asked to assess sentence relevance to a particular domain in the Risk of Bias tool using a Likert-like scale (3=highly relevant, 2=somewhat relevant, and 1=not relevant). The assessors were provided with additional definitions for each of these categories. Two assessors piloted the evaluation and had substantial agreement (Kappa=0.79). We calculated based on pilot data that at least 350 trials would be needed to detect a 10% difference in model output quality with 80% power with significance of $P < 0.05$. We collected a total of 1731 judgments from 20 experts from 371 trials.

Results

Document level results

Model 2 judgments were less accurate than those from published reviews, though the difference was <10% (overall accuracy 71.0% with ML *v* 78.3% with CDSR; P<0.001). Model 1 (which does not include supporting sentences and models each domain separately) achieved substantially greater accuracy than baseline. Model 2 (which jointly models all domains, and incorporates information about whether sentences are judged relevant) improved performance compared with Model 1 in all but one domain (*blinding of outcome assessment*), though our study was not powered to assess the statistical significance of this difference. Model 3, which ignores the noisy *selective reporting* and *incomplete reporting of outcomes* domains, resulted in uniform improvement compared to Model 1 across all the domains it included.

Table 1 Results from the document evaluation task: Baseline=accuracy achieved by labeling all test documents with majority class for that domain; Model 1=separate bag-of-words model for each domain; Model 2=multi-task model jointly modeling all domains and incorporating information about sentence relevance as features; Model 3=same multi-task model excluding domains 5 and 6; Cochrane=estimate of human accuracy obtained by comparing a second risk of bias assessment (of the same trials) from another systematic review

Domain	Trials (n)	Baseline	Model 1	Model 2	Model 3	Cochrane	P (Model 2 versus Cochrane)
Overall	6610	56.4%	69.3%	71.0%	-	78.3%	P < 0.001
1. Random sequence generation	1225	59.3%	72.5%	73.9%	75.8%	84.8%	P < 0.001
2. Allocation concealment	2089	53.7%	72.4%	74.0%	73.3%	80.0%	P < 0.001
3. Blinding of participants and personnel	1051	50.4%	72.6%	73.0%	73.7%	78.1%	P = 0.003
4. Blinding of outcome assessment	250	57.7%	64.0%	61.5%	67.4%	83.2%	P < 0.001
5. Incomplete reporting of outcomes	1149	60.9%	63.9%	65.1%	-	71.3%	P < 0.001
6. Selective reporting	846	59.9%	61.8%	67.6%	-	73.0%	P = 0.010

Sentence level results

The results from task 2 (identifying sentence with information about risk of bias) are presented in tables 2 and 3. The *top1* model (which retrieves one sentence per study) performed substantially better than baseline, but produced text judged less relevant than that in the CDSR (10% fewer documents with highly relevant output, and 14% fewer documents with highly, or somewhat relevant output). The best text from the *top3* model was rated as more relevant than text from the CDSR overall, and in individual domains, but differences were not statistically significant.

Table 2 Proportion of studies for which highly relevant text was identified using four methods: baseline, one random sentence chosen per document; top-1, the one most informative sentence according to the algorithm; top-3, the top three most informative sentences according to the algorithm; and cochrane, being text quoted in published Cochrane reviews to justify bias decisions (mean 1.3 sentences per document). Where more than one sentence was identified, the one highest rated sentence contributes to the score.

Domain	Trials (n)	baseline	top1	top3	cochrane	top1 v cochrane	top3 v cochrane
Overall	378	0.5%	45.0%	60.4%	56.5%	-11.6% (-18.5% to -4.4%); P<0.001	+3.9%, (-3.2% to +10.9%); P=0.141
1. Random sequence generation	81	0.0%	55.6%	65.4%	60.5%		
2. Allocation concealment	75	0.0%	44.0%	60.0%	60.0%		
3. Blinding of participants and personnel	76	0.0%	55.3%	72.4%	68.4%		
4. Blinding of outcome assessment	56	0.0%	39.3%	62.5%	57.1%		
5. Incomplete reporting of outcomes	67	3.0%	40.9%	57.6%	50.8%		
6. Selective reporting	23	0.0%	0.0%	4.6%	4.6%		

Table 3 Proportion of studies for which highly or somewhat relevant text was identified using four methods; see caption of table 2 for details of models

Domain	Trials (n)	baseline	top1	top3	cochrane	top1 v cochrane	top3 v cochrane
Overall	378	3.7%	69.7%	84.9%	83.8%	-14.1% (-20.0% to -8.1%); P<0.001	+1.0% (-4.2% to +6.2%); P=0.35
1. Random sequence generation	81	4.9%	88.9%	92.6%	88.9%		
2. Allocation concealment	75	0.0%	72.0%	88.0%	89.3%		
3. Blinding of participants and personnel	75	4.0%	68.5%	84.2%	81.6%		
4. Blinding of outcome assessment	56	3.6%	58.9%	83.9%	82.1%		
5. Incomplete reporting of outcomes	67	7.5%	71.2%	90.9%	88.1%		
6. Selective reporting	23	0.0%	18.2%	31.9%	45.5%		

Discussion

In this paper we reported the development and evaluation of RobotReviewer, a system for automating Risk of Bias assessment. Our system determines whether a trial is at low risk of bias for each domain in the Cochrane Risk of Bias tool, and it identifies text that supports the judgment. We demonstrated strong performance on these tasks. Automatic document judgments were of reasonable accuracy, lagging our estimate of human reviewer accuracy by around 7%. Our automated approach identified text supporting risk of bias judgments of similar quality to that found in published systematic reviews.

While our algorithm is not ready to replace manual risk of bias assessment altogether, we envisage several ways it could reduce author workload in practice. Since justifications are provided, most errors should be easy to identify, meaning reviewers need consult the full paper only where judgments are not adequately justified (see Box 3). This should alleviate a common concern about automation technologies: that they act as *black boxes*.(18) Elsewhere, we have described a prototype tool which presents the model predictions to the user directly within the original PDF document.(19) Figure 5 shows the system in use. This has the additional advantage of preserving the link between published reviews and their source data, a key omission in current practice.(20) Alternatively, this system could also be used to draw reviewers' attention to sentences that are likely to be relevant, leaving them to make the final judgment: this would speed risk of bias assessment. Finally, we note that current practice is for two reviewers to assess the risk of bias of trials independently, then reach a consensus. An alternative workflow would be to replace one of these two reviewers with the automated approach, thus still having a second independent assessment.

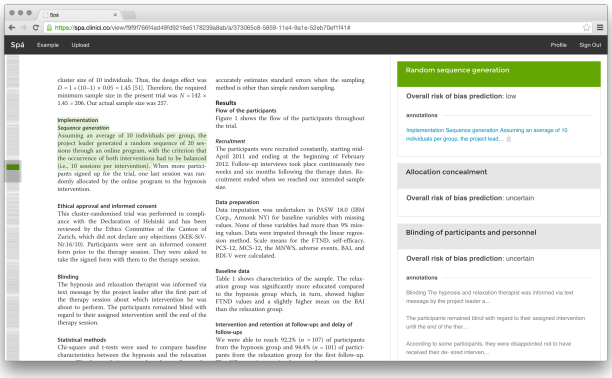

Figure 5: Example of our prototype system showing the bias assessment for random sequence generation and a supporting sentence.

One potential weakness of our approach is that our corpus comprises a fraction of the studies reported in the Cochrane Library. The 12,808 PDFs were a convenience sample comprised of PDFs available via university library subscriptions plus open access publications. Studies without an associated PDF comprise those available only in HTML or paper formats, and those for which our institutions did not hold a subscription. Studies with unobtainable PDFs might be expected to have increased risks of bias, particularly those in lesser-known journals, those reported in conference abstracts only, and those (typically older) study reports available in paper format only. We found in a post-hoc analysis that these were 6% less likely to be at low risk of bias on average. However, our system is designed to be used on PDFs, and our corpus should therefore be representative of (and thus generalize to) the types of PDFs which researchers use.

In this work we sought to estimate the accuracy of manual risk of bias assessment by making use of trials with 2 or more RoB assessments in the CDSR. Our approach makes the simplifying assumption that all discrepancies between two Risk of Bias assessments represent errors or subjective disagreements. In practice, this will not always be the case, and risk of bias judgments may be influenced by individual review question, or by outcome (for some bias types). However, we note that our estimate of human performance shows substantially greater agreement for this task than previous estimates. In practice, where multiple Cochrane reviews contain the same trial, it is likely that they were produced by the same review group, and will share editors, and possibly author teams who may reach similar bias decisions more often than expected by independent groups.

Domain	Random sequence generation
Risk of bias	Low
Text justifying judgment	*Sequence generation* *Assuming an average of 10 individuals per group, the project leader generated a random sequence of 20 sessions through an online program, with the criterion that the occurrence of both interventions had to be balanced (i.e., 10 sessions per intervention)*

Box 3 Example of model output where a reviewer could verify the judgment by reference to the justifying text, without reference to the original paper.

There are several promising routes for improving model performance. First, our current model was trained on trials from any clinical specialty. Since reviewer agreement increases when review specific guidance is given,(5) training a model on trials relevant to a review of interest may improve performance.

Second, the Cochrane Handbook recommends that some domains of bias should be assessed per *outcome*, rather than per study.(15) While our tool should identify text that is relevant to bias for all outcomes assessed, it produces one overall judgment per paper. This may partly explain relatively poor performance in the domain *blinding of outcome assessment*, which seems likely to vary substantially for different outcomes in the same trial. This is the approach taken by the vast majority of Cochrane reviews at present also (which form our training data), but we aspire to judge bias for each outcome in future versions of our tool.

Third, the domain *selective outcome reporting* requires reference to a trial protocol, to find if any pre-specified outcomes were not reported in the final paper. The fact that our model performed better than baseline for this domain is probably explained by the fact that selecting reporting bias is correlated with other biases that are more readily determined from the trial publication. Compulsory registration of clinical trials means that trial protocols are increasingly easy to obtain; indeed, the WHO collates protocols from international registries in machine-readable format, including lists of outcomes. Linking with this dataset may be yield better performance.

And finally, by incorporating reviewer corrections back into the model in real time, we could make use of *online* learning. Such a strategy would, in theory, improve model performance with use. This would make it possible to stay up-to-date with changes in research methodology or bias assessment practice over time, and additionally would learn to assess bias in a way which is tailored to the needs of a particular clinical area, review group, or even individual authors.

Conclusion

We have outlined a method for automating risk of bias assessment, including the tasks of both categorizing articles as at *low* or *high/unclear* risk and extraction of sentences supporting these categorizations, across several domains. While model performance lags behind human accuracy, it can identify text in trial PDFs with similar relevance to that used in published reviews. Future methodological improvements are likely to close the gap between the algorithm and human performance, and may eventually be able to replace manual risk of bias assessment altogether.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing Interests Statement

The authors have no competing interests to declare.

Contributorship Statement

All authors contributed to the study concept, data linkage, corpus construction, model development, validation study design, recruitment of the expert panel, and data analysis. All authors jointly drafted the manuscript, and all have approved the final version of the manuscript. IM is the corresponding author.

Acknowledgments We would like to thank the following people who participated in our expert panel:

- Tom Trikalinos, Director of the Centre of Evidence Based Medicine (CEBM), Brown University
- Rachel Marshall, Editor, Cochrane Editorial Unit
- Emma Welsh, Managing Editor, Cochrane Airways Group
- Kayleigh Kew, Systematic reviewer, Cochrane Airways Group
- Ethan Balk, Assistant professor, CEBM, Brown University
- James Thomas, Professor, Institute of Education, University of London
- Michelle Richardson, Researcher, Institute of Education, University of London
- Mark Newman, Reader in evidence-based policy, Institute of Education, University of London
- Oluwaseun Akinyede, Research assistant, ECRI institute
- Jonathan Treadwell, Associate director, ECRI institute
- Joann Fontanarosa, Research assistant, ECRI institute
- Karla Soares-Weiser, Managing Director, Enhance Reviews
- Hanna Bergman, Senior researcher, Enhance Reviews
- Mei Chung, Professor of Evidence Based Medicine, Tufts University
- Issa Dahabreh, Assistant professor, CEBM, Brown University
- Matthew Page, Systematic reviewer, Monash University
- Miranda Cumpston, Editor, Cochrane Public Health Group
- Ian Shemilt, Senior Research Associate, University of Cambridge
- Jani Ruotsalainen, Managing Editor, Cochrane Occupational Safety and Health Review Group
- Therese Dalsbø, Senior Adviser, Norwegian Knowledge Centre for the Health Services

We thank Chris Mavergames at the Cochrane Collaboration for his help in recruiting the panel.

References

1. Centre for Reviews and Dissemination. Assessing quality. In Systematic reviews: CRD's guidance for undertaking reviews in healthcare. CRD, University of York; 2009. p. 33.

2. Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ [Internet]. 2011 Oct;343(oct18 2):d5928. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.d5928

3. Hartling L, Ospina M, Liang Y. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ [Internet]. 2009;339:b4012. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2764034/

4. Tovey D, Marshall R, Hopewell S, Rader T. Fir for purpose: centralising updating support for high-priority Cochrane Reviews. National Institue for Health Research Evaluation, Trials,; Studies Coordinating Centre; 2011 pp. 1–66. Report No.: July.

5. Hartling L, Bond K, Vandermeer B, Seida J, Dryden DM, Rowe BH. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. PloS one [Internet]. 2011 Jan;6(2):e17242. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3044729\&tool=pmcentrez\&rendertype=abstract

6. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Systematic reviews [Internet]. 2014 Jan;3(1):74. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4100748\&tool=pmcentrez\&rendertype=abstract

7. Marshall I, Kuiper J, Wallace B. Automating Risk of Bias Assessment for Clinical Trials. In Proceedings of the aCM conference on bioinformatics, computational biology and biomedicine [Internet]. 2014. Available from: http://www.cebm.brown.edu/static/papers/acmbcb2014_final.pdf

8. Summerscales RL. Automatic Summarization of Clinical Abstracts for Evidence-Based Medicine [Internet] [PhD thesis]. Illinois Institute of Technology; 2013. Available from: http://www.andrews.edu/~summersc/summerscales_phdthesis2013.pdf

9. Kiritchenko S, Bruijn B de, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Medical Informatics and Decision Making [Internet]. 2010 Sep;10(1):56. Available from: http://www.biomedcentral.com/1472-6947/10/56/abstract http://www.biomedcentral.com/1472-6947/10/56/

10. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In [Internet]. Suntec, Singapore: Association for Computational Linguistics; 2009. pp. 1003–11. Available from: http://www.aclweb.org/anthology/P/P09/P09-1113

11. Ling, X, Jiang, J, He, X, Mei, Q, Zhai, C, & Schatz, B. Automatically generating gene summaries from biomedical literature; Pacific Symposium on Biocomputing (PSB) 2006.

12. Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. In Proceedings of the 39th annual meeting on association for computational linguistics [Internet]. 2001. pp. 26–33. Available from: http://dl.acm.org/citation.cfm?id=1073017

13. xPDF, Open Source PDF viewer, http://www.foolabs.com/xpdf/ [Accessed January 2015]

14. Caruana, Rich. Multitask learning. Springer US, 1998.

15. Higgins J, Altman D, Sterna J. Assessing risk of bias in included studies. In Cochrane handbook for systematic reviews of interventions version 510 (updated march 2011) [Internet]. 2011. Available from: www.cochrane-handbook.org

16. Daumé H. Frustratingly easy domain adaptation. In DANLP 2010 proceedings of the 2010 workshop on domain adaptation for natural language processing [Internet]. 2010. Available from: http://arxiv.org/abs/0907.1815

17. Weinberger K, Dasgupta A, Langford J, Smola A, Attenberg J. Feature hashing for large scale multitask learning. In Proceedings of the 26th annual international conference on machine learning. ACM; 2009. pp. 1113–20.

18. Thomas J. Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? OA Evidence-Based Medicine [Internet]. 2013 Oct;1(2):1–6. Available from: http://www.oapublishinglondon.com/article/1109

19. Kuiper J, Marshall I, Wallace B, Swertz M. Spá: A web-based viewer for text mining in evidence based medicine. In: Calders T, Esposito F, Hüllermeier E, Meo R, editors. Machine learning and knowledge discovery in databases [Internet]. Springer Berlin Heidelberg; 2014. pp. 452–5. Available from: http://dx.doi.org/10.1007/978-3-662-44845-8_33

20. Adams CE, Polzmacher S, Wolff A. Systematic reviews: work that needs to be done and not to be done. Journal of evidence-based medicine [Internet]. 2013 Nov;6(4):232–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24325416